One Versus all for deep Neural Network Incertitude (OVNNI) quantification

Gianni Franchi[1], Andrei Bursuc[2], Emanuel Aldea[3], Séverine Dubuisson[4] and Isabelle Bloch[5]
1: ENSTA Paris, Institut polytechnique de Paris, 2: valeo.ai, 3: SATIE, Université Paris-Saclay, 4: CNRS, LIS, Aix Marseille University, 5: Télécom Paris, Institut polytechnique de Paris

Context and objectives

Deep neural networks (DNNs) are powerful learning models yet their results are not always reliable.

- In this work we aim for efficient deep DNNs able to quantify the epistemic uncertainty of data easily.
- We achieve this task by training multiple One vs All DNNS and one All vs All DNN.
- Our approach achieves state of the art performance in quantifying OOD data across multiple datasets and architectures while requiring little hyper-parameter tuning.

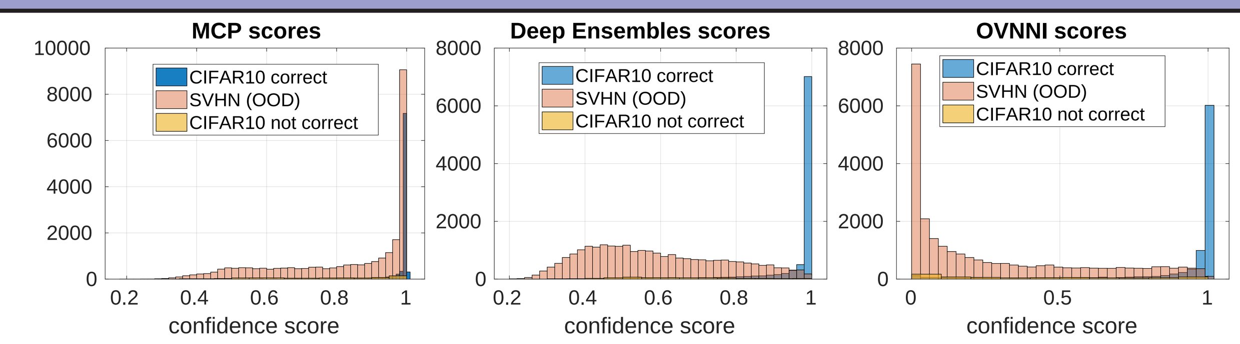

Fig. 1: Distribution of classifications scores. we have respresented the histograms of confidence scores of Resnet50 [He+16] trained on the CIFAR10 [KH+09] training set and tested on SVHN [Net+11] and CIFAR10 testing set, using Maximum Class Probability (MCP) [HG16], Deep Ensembles [LPB17], and OVNNI,

Deep Neural Network (DNN) and epistemic uncertainty

- **BNNs** [Blu+15]: aim to find the posterior distribution of the parameters given the training dataset $P(\Theta \mid \mathcal{D})$, not only the values corresponding to the MAP. To make a prediction y on a new sample $\mathbf{x}$ the BNN compute : $P(y \mid \mathbf{x}, \mathcal{D}) = \int P(y \mid \mathbf{x}, \Theta) P(\Theta \mid \mathcal{D}) d\Theta$.
- **Deep Ensembles**[LPB17]: train multiple DNNs to have access to their uncertainty.
- **One vs All (OVA)/ One vs One (OVO) ensembles**: popular techniques for performing multi-label classification based on an ensemble of binary base classifiers. For OVO, instead of the baseline max-voting aggregation strategy, pairwise coupling [WLW04] or ECOC [DB94] have been widely used. Recently a new approach [Pad+20] mixing OVA and deep learning had interesting results.

From AVA to OVA

Classically we use Cross entropy defined on a batch B of size $N \in \mathbb{N}$ by:

$$\mathcal{L}(\boldsymbol{\omega}(t), B) = -\frac{1}{N} \sum_{i=1}^{N} \log(P(Y = y_i \mid X = x_i, \boldsymbol{\omega})) \quad (1)$$

We train one OVA DNN of each class j that provides $P(Y_j = 1 \mid X = x_i, \boldsymbol{\omega}^j)$, and one AVA DNN that provides $P(Y = j \mid X = x_i, \boldsymbol{\omega})$ for all j in $[1, n_{\text{label}}]$. We consider that the final confidence score for a data x_i to belong to class j is:

$$p_j(x_i) = P(Y_j = 1 \mid X = x_i, \boldsymbol{\omega}^j) \times P(Y = j \mid X = x_i, \boldsymbol{\omega}) \quad (2)$$

OVNNI

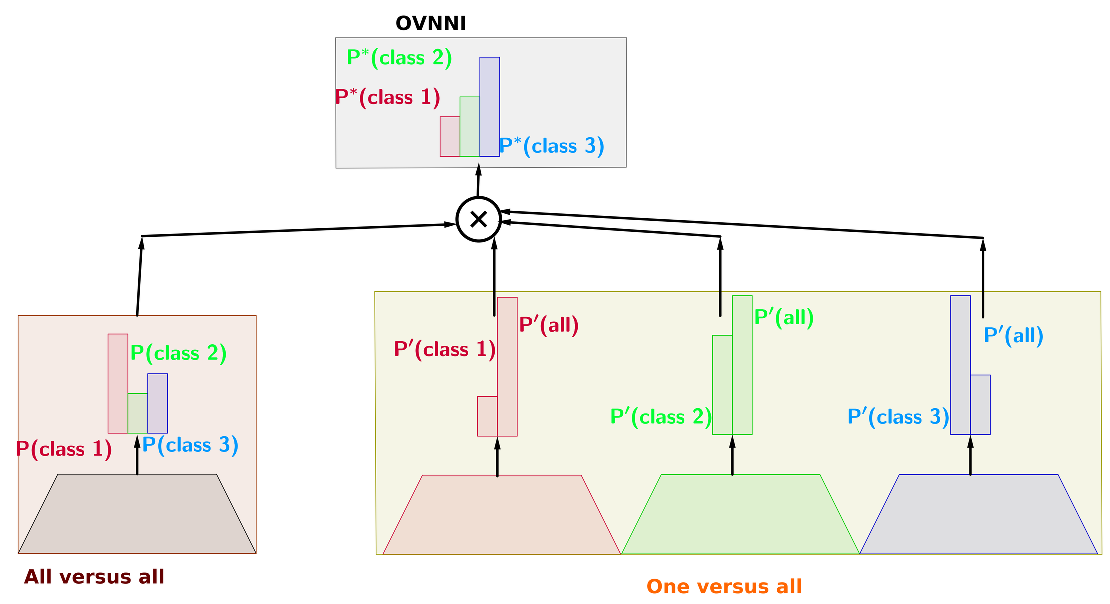

Fig. 2: From AVA and OVA to OVNNI process in the case we deal with a database composed of just three classes.

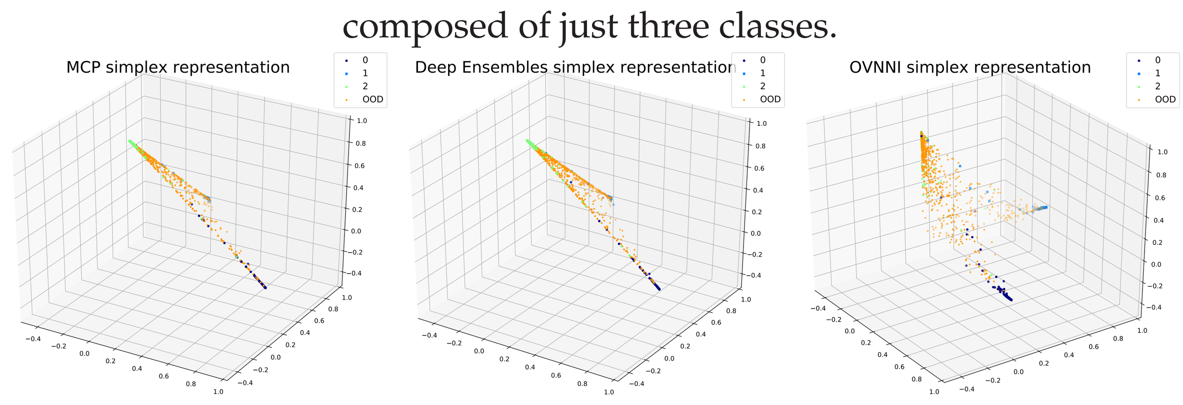

Fig. 3: Results on MNIST - 3 classes experiments. We represent in these figures the softmax prediction outputs obtained by the baselines (a) MCP, (b) Deep Ensemble, and (c) by OVNNI, respectively.

Experimental results

We evaluate the performance of LP-BNN in assessing the uncertainty of its predictions on **MNIST**[LeC+98], **CIFAR-10** [KH+09] **StreetHazards** [Hen+19], and **BDD-Anomaly**[Hen+19].

Dataset	OOD technique	Accuracy/mIoU	AUC	AUPR Error	AUPR Success	ECE	Real ECE
MNIST/Not MNIST 3 hidden layers	Baseline (MCP)	98.8	92.7	96.1	81.4	0.305	
	MCP + One class SVM	98.8	97.4	98.4	95.9	0.072	
	MC Dropout	98.2	88.1	89.8	81.7	0.494	
	Deep Ensemble	**98.9**	97.7	98.4	95.8	0.462	
	TRADI	98.6	97.1	98.4	94.6	0.407	
	ODIN	98.8	94.2	96.8	85.6	0.500	
	ConfidNET	98.2	97.4	98.8	94.1	0.461	
	Ensemble OVA (ours)	97.2	**99.0**	**99.5**	**97.3**	0.179	
	OVNNI (ours)	98.8	**99.1**	**99.6**	**97.9**	0.066	
CIFAR10 ResNet50	Baseline (MCP)	93.1	83.9	92.9	67.5	0.606	0.0278
	MCP +One class SVM	93.1	79.7	90.9	63.5	0.203	0.5881
	MC Dropout	93.1	83.9	92.9	67.5	0.606	0.0278
	Deep Ensemble	**95.0**	**95.8**	**97.7**	**92.1**	0.422	0.0124
	ODIN	93.1	83.9	93.3	67.2	0.606	0.0483
	ConfidNET	93.1	85.1	94.6	61.2	0.706	0.0529
	Ensemble OVA (ours)	89.3	91.8	95.8	87.1	0.468	0.0803
	OVNNI (ours)	93.3	94.3	**97.3**	**91.1**	**0.187**	0.0185
Camvid ENET	Baseline (MCP)	85.8/52.9	79.7	52.1	92.6	0.146	
	MC Dropout	80.3/48.6	80.2	56.1	89.3	0.168	
	Deep Ensemble	88.0/58.2	83.2	54.3	94.0	0.112	
	TRADI	83.4/51.4	83.2	55.9	93.8	0.110	
	ConfidNET	83.4/52.8	81.3	58.3	92.6	0.121	
	Ensemble OVA (ours)	87.9/52.8	91.7	69.6	98.4	0.060	
	OVNNI (ours)	93.1/66.1	**94.0**	**75.7**	**99.0**	0.025	
StreetHazards PSPNet (ResNet50)	Baseline (MCP)	90.0/54.6	91.6	50.8	98.9	0.055	
	MC Dropout	88.0/47.9	88.8	51.8	97.8	0.092	
	Deep Ensemble	90.2/55.0	92.2	52.0	99.0	0.051	
	TRADI	90.2 /54.6	92.1	51.4	99.1	0.049	
	ConfidNET	90.0/54.6	88.9	37.0	97.9	0.10	
	Ensemble OVA (ours)	89.7/54.0	92.4	52.3	**99.1**	**0.048**	
	OVNNI (ours)	90.0/54.6	**93.0**	**53.4**	**99.2**	**0.048**	
BDD Anomaly PSPNet (ResNet50)	Baseline (MCP)	89.9/52.8	81.4	62.5	91.5	0.159	
	MC Dropout	88.7/49.5	76.0	55.7	88.2	0.181	
	Deep Ensemble	91.0/57.6	85.5	67.3	93.9	0.170	
	TRADI	89.9/52.1	81.9	63.2	91.8	0.157	
	ConfidNET	89.9/52.8	78.3	56.4	91.2	0.232	
	Ensemble OVA (ours)	89.9/52.8	**91.2**	86.2	95.7	**0.072**	
	OVNNI (ours)	90.7/55.4	**91.9**	**86.6**	**95.9**	**0.081**	

Tab. 1: Comparative results for classification tasks on CIFAR-10 and CIFAR-100. The results are averaged over three seeds.

Dataset	OOD technique	AUC	AUPR	FPR-95%-TPR
MNIST/Not MNIST 3 hidden layers	Baseline (MCP)	94.0	96.0	24.6
	MCP + One class SVM	96.9	98.0	12.5
	MC Dropout	91.8	94.9	35.6
	Deep Ensemble	97.2	98.0	9.2
	TRADI	96.7	97.6	11.0
	ODIN	94.9	96.7	17.5
	ConfidNET	97.9	99.0	12.7
	Ensemble OVA (ours)	98.9	**99.4**	5.9
	OVNNI (ours)	**99.3**	**99.6**	**3.5**
CIFAR10 ResNet50	Baseline (MCP)	80.4	89.7	61.5
	MCP + One class SVM	78.8	89.6	61.5
	MC Dropout	80.4	89.7	62.6
	Deep Ensemble	**93.0**	**96.2**	**19.3**
	ODIN	80.3	89.9	61.3
	ConfidNET	84.8	94.0	68.3
	Ensemble OVA (ours)	88.5	93.0	30.9
	OVNNI (ours)	92.2	**95.8**	23.3
Camvid ENET	Baseline (MCP)	75.4	10.0	65.1
	MC Dropout	75.4	10.7	63.2
	Deep Ensemble	79.7	13.0	55.3
	TRADI	79.3	12.8	57.7
	ConfidNET	81.9	13.8	55.8
	Ensemble OVA (ours)	**97.1**	**71.1**	**13.5**
	OVNNI (ours)	**96.1**	**61.2**	**16.5**
StreetHazards PSPNet (ResNet50)	Baseline (MCP)	88.7	6.9	26.9
	MC Dropout	69.9	6.0	32.0
	Deep Ensemble	90.0	7.2	25.4
	TRADI	89.2	7.2	25.3
	ConfidNET	83.6	2.3	26.2
	Ensemble OVA (ours)	**91.6**	**12.7**	**21.9**
	OVNNI (ours)	**91.2**	**12.6**	**22.2**
BDD Anomaly PSPNet (ResNet50)	Baseline (MCP)	86.0	5.4	27.7
	MC Dropout	85.2	5.0	29.3
	Deep Ensemble	87.0	6.0	25.0
	TRADI	86.1	5.6	26.9
	ConfidNET	85.4	5.1	29.1
	Ensemble OVA (ours)	87.0	4.9	29.0
	OVNNI (ours)	**87.2**	**6.7**	**25.0**

Tab. 2: Comparative results obtained on the OOD task for semantic segmentation. The results are averaged over three seeds.

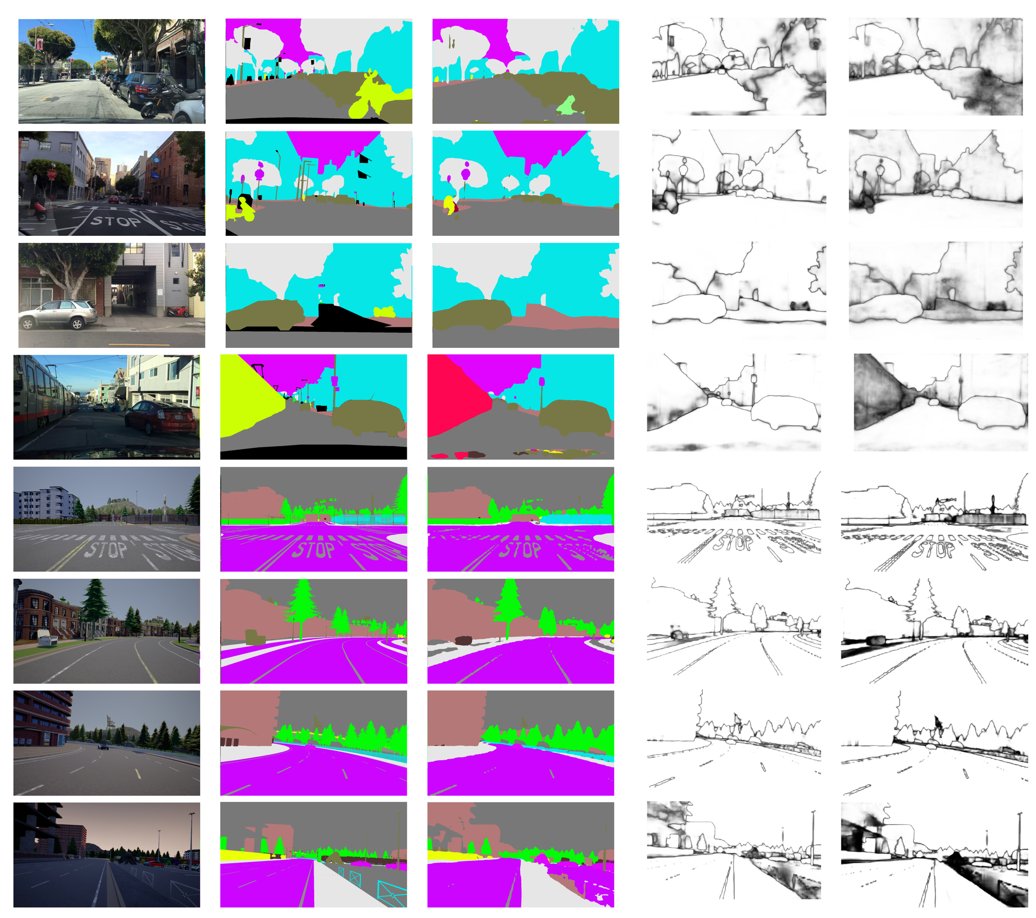

Fig. 4: Results of OVNNI on BDD Anomaly and StreetHazards. The first column is the input image, the second is the ground truth, the third is prediction and the fifth is the confidence score of OVNNI. For comparison, we add the MCP confidence score in the fourth column. We can see that OVNNI has a low score on the motorcycle on the three first rows and on the train on the last row which correspond to the OOD classes.

Conclusions

In this work, we presented an approach based on One versus All training and mixed with a modern approach based on deep learning. We show that the combination of these approaches reaches states of the art performance on all segmentation experiments. Regarding classification tasks, OVNNI exhibits the best calibration performance. Concurrent approaches suffer from a lack of performance in calibration in most datasets, hence the scores that they provide are overconfident, potentially leading to dangerous scenarios in critical applications. In addition to the reported performance, our approach needs little hyperparameter tuning and is easy to implement.

References

[Blu+15] Charles Blundell et al. "Weight uncertainty in neural networks". In: *arXiv preprint arXiv:1505.05424* (2015).

[DB94] Thomas G Dietterich and Ghulum Bakiri. "Solving multiclass learning problems via error-correcting output codes". In: *Journal of artificial intelligence research* 2 (1994), pp. 263–286.

[He+16] Kaiming He et al. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[Hen+19] Dan Hendrycks et al. "A Benchmark for Anomaly Segmentation". In: *arXiv preprint arXiv:1911.11132* (2019).

[HG16] Dan Hendrycks and Kevin Gimpel. "A baseline for detecting misclassified and out-of-distribution examples in neural networks". In: *arXiv preprint arXiv:1610.02136* (2016).

[KH+09] Alex Krizhevsky, Geoffrey Hinton, et al. *Learning multiple layers of features from tiny images*. Tech. rep. Citeseer, 2009.

[LeC+98] Yann LeCun et al. "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324.

[LPB17] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. "Simple and scalable predictive uncertainty estimation using deep ensembles". In: *Advances in Neural Information Processing Systems*. 2017, pp. 6402–6413.

[Net+11] Yuval Netzer et al. "Reading Digits in Natural Images with Unsupervised Feature Learning". In: (2011).

[Pad+20] Shreyas Padhy et al. "Revisiting One-vs-All Classifiers for Predictive Uncertainty and Out-of-Distribution Detection in Neural Networks". In: *arXiv preprint arXiv:2007.05134* (2020).

[WLW04] Ting-Fan Wu, Chih-Jen Lin, and Ruby C Weng. "Probability estimates for multi-class classification by pairwise coupling". In: *Journal of Machine Learning Research* 5.Aug (2004), pp. 975–1005.